VISION PAPER

Not Ready for Convergence in Data Infrastructures

Keith Jeffery[1†], Peter Wittenburg[2], Larry Lannom[3], George Strawn[4], Claudia Biniossek[5], Dirk Betz[5] & Christophe Blanchi[6]

[1]Keith G Jeffery Consultants, 71 Gilligans Way, Faringdon SN7 7FX, UK

[2]Max Planck Computing and Data Facility, Gießenbachstraße 2, 85748 Garching, Germany

[3]Corporation for National Research Initiatives (CNRI), Reston, Virginia 20191, USA

[4]US National Academy of Sciences, Washington DC 20418, USA

[5]GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667, Cologne, Germany

[6]Corporation for National Research Initiatives, Reston, Virginia 20191, USA

Keywords: Scientific process; Workflow; Metadata; FAIR; Scientific data; Data wrangling

Citation: Jeffery, K., et al.: Not ready for convergence in data infrastructures. Data Intelligence 3(1), 116-135 (2021). doi: 10.1162/dint_a_00084

ABSTRACT

Much research is dependent on Information and Communication Technologies (ICT). Researchers in different research domains have set up their own ICT systems (data labs) to support their research, from data collection (observation, experiment, simulation) through analysis (analytics, visualisation) to publication. However, too frequently the Digital Objects (DOs) upon which the research results are based are not curated and thus neither available for reproduction of the research nor utilization for other (e.g., multidisciplinary) research purposes. The key to curation is rich metadata recording not only a description of the DO and the conditions of its use but also the provenance – the trail of actions performed on the DO along the research workflow. There are increasing real-world requirements for multidisciplinary research. With DOs in domain-specific ICT systems (silos), commonly with inadequate metadata, such research is hindered. Despite wide agreement on principles for achieving FAIR (findable, accessible, interoperable, and reusable) utilization of research data, current practices fall short. FAIR DOs offer a way forward. The paradoxes, barriers and possible solutions are examined. The key is persuading the researcher to adopt best practices which implies decreasing the cost (easy to use autonomic tools) and increasing the benefit (incentives such as acknowledgement and citation) while maintaining researcher independence and flexibility.

[†] Corresponding author: Keith Jeffery (Email: keith.jeffery@keithgjefferyconsultants.co.uk; ORCID: 0000-0003-4053-7825).

1. INTRODUCTION

In the paper *Common Patterns in Revolutionary Infrastructures and Data* [1] Wittenburg and Strawn compared the phases of the evolution of some fundamental infrastructures (electricity, Internet, and Web) as historical examples and related them to the evolution of a "data infrastructure based on commons" which will be revolutionary and disruptive as well. It turns out that the emergence of all these "revolutionary" infrastructures follows similar patterns from early visions up to an intensive exploitation phase (Figure 1). They gave hope that basic ingredients such as (1) the FAIR principles [2], (2) Digital Objects (DOs) as originally defined by Kahn and Wilensky [3], extended by Research Data Alliance (RDA) based on many use cases[①] [4] and finally extended to FAIR DOs as defined at a recent Paris workshop [5] and (3) exemplary work on the syntax and semantics of data and metadata could be seen as roots for convergence to build a common and efficient data infrastructure [6]. Some recent observations and developments from a pilot survey from GEDE/GO FAIR [7] and a detailed analysis of more than 50 research infrastructure initiatives, however, made clear that convergence enabling a vigorous data exploitation phase might be further away than was hoped due to the continuing huge inefficiencies they found (about 80% of the effort in data driven projects is wasted with data wrangling). This paper proposes approaches to reaching the convergence, with the hope of speeding up the transition towards an intensive data exploitation phase.

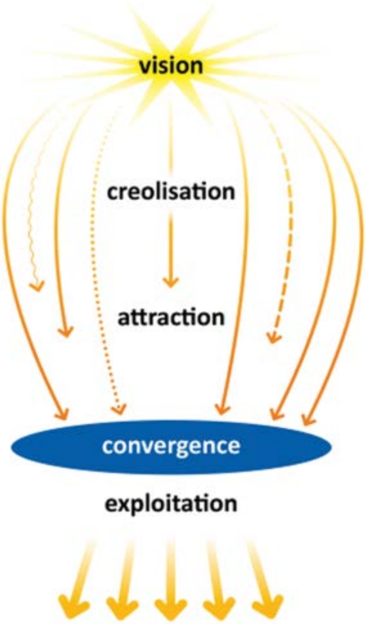

Figure 1. The phases necessary to come from a vision to an intensive exploitation phase.

[①] A DO is an abstract concept that has a content represented by a structured bit-sequence which is stored in some trustworthy repositories, has associated metadata and is assigned a globally unique, persistent and resolvable identifier (PID).

It should be noted that the topics we are discussing in this paper are relevant for data driven work that is being carried out crossing silo and/or disciplinary borders. Since most researchers still work on their own data sets or work within narrow project boundaries, they could, in principle, define their own private rules and "standards" to optimize efficiency. However, practices in large projects, for example, in cancer or language research, have shown that individual researchers or small focussed research teams also have severe problems in remembering after a few months what they exactly did when the data volumes and complexity[2] increased to or beyond a certain level. This is especially true when there is no obvious regular mechanism that inherently structures the created data and where the traditional file mechanisms (naming, directory structure, etc.) are not sufficient anymore. Furthermore, researchers working in silos may suddenly find that their digital materials are, indeed, of use in multidisciplinary studies and so best practices should have been followed to enable reuse.

In this paper we want to first describe the practices in the data labs[3], second describe some approaches and guidelines that could help to overcome the inefficiencies, third present some urgent steps that should be taken, and finally draw a number of conclusions.

2. SITUATION IN THE DATA LABS

Based on a first broad survey in 2014 [8], the RDA Data Fabric Interest Group (DFIG) was set up and had as its focus the improvement of the situation in the data labs [9]. As indicated in the abstract diagram (Figure 2) the situation in the labs was distinguished from the data publication step which can be the final act of data driven work. For the experts participating in the DFIG it was obvious that most data will exist and be processed in this fundamental cycle, but only limited data (<10 %) will be occasionally published in the "classical publication sense", often in form of collections associated with some scientific paper[4]. Efficient Data Science (DS) will depend on the early and stable availability of data and other artifacts which require the following simple (but obviously difficult to achieve) measures: (1) They need to be stored, managed and made accessible by trustworthy repositories for long periods of time. (2) They need to be identified by universally unique, persistent and resolvable identifiers. (3) They need to be associated with suitable and comprehensive metadata. These three pillars make up what the concept of (FAIR) DOs basically implies.

[2] With "complexity" we do not just indicate increasing volumes and heterogeneity of digital objects (DOs), but also include many different types of relations between those DOs.

[3] The term "data labs" includes all laboratories, departments and projects which have a focus on creating, managing, curating and processing data as part of experimenting, observing, simulating and computing activities.

[4] In most cases all process data (context and provenance) is not disseminated reducing re-usability.

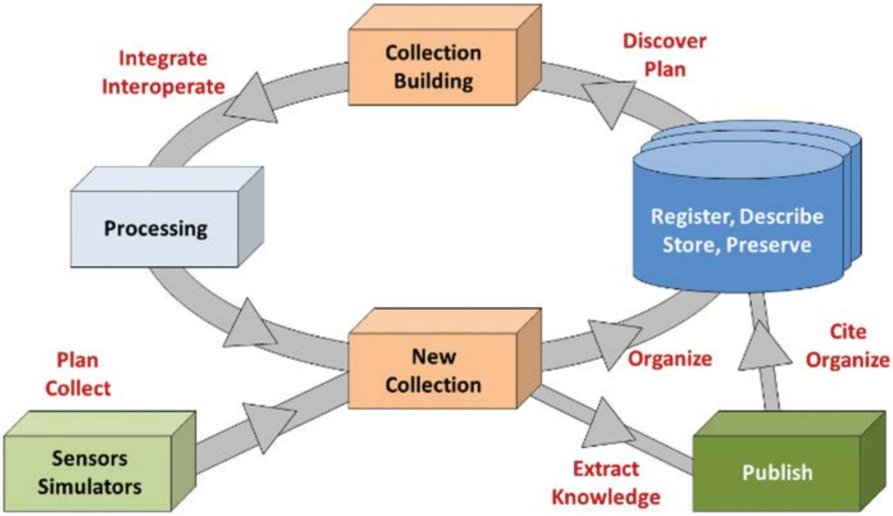

Figure 2. A key diagram used in the RDA Data Fabric IG. Note: It characterizes the continuous work in labs focusing on data-driven science. Raw data are organized and stored in trustworthy repositories together with all sorts of derived data. Collections are built based on scientific insights and made subject of processing to compute new derived data also being stored in repositories. In some cases, data publications will occur as final steps of this continuous cycle.

In the following we will describe a few major observations which we can extract from a recent pilot survey from GEDE/GO FAIR [7] and a deep analysis of more than 50 research infrastructure initiatives.

2.1 State of Awareness and Coherence

In general, the awareness of data matters has increased since almost all researchers are facing similar challenges due to increasing volumes, heterogeneity (types, formats, and semantics)[⑤] and complexity which includes the extensive relationships between DOs. However, we can observe different states of awareness about the requirements for modern data-driven science between researchers, research disciplines and countries. Improved awareness is often paralleled with a higher degree of coherence in views and terminology and infrastructural measures at the national or organizational level. While the agreements and investments with respect to "hard infrastructural matters" (networks, HPC, storage, and physical clouds) are high, the agreements on "soft matters" which are characterized by the scope of the FAIR principles are behind, despite all claims on paper. There are many different interpretations of the FAIR principles and different views lead to a large variety of suggestions hampering progress.

It is a widely used practice now to claim support for Openness and FAIRness. When analyzing the practices we can observe (1) that mostly only the findability and accessibility dimensions are addressed

[⑤] We can refer here to the often used 5Vs (volume, velocity, variety, veracity, value).

and that "machine actionability"[6] which is in the core of FAIR is not well understood, (2) that following the principles is often seen as an administrative add-on and not yet as a step to enable new science or even to improve the usefulness through additional information of existing science and (3) that the interpretation scope is limited as the following elaboration indicates. Researchers in almost all fields are using sensors that create increasing amounts of data and are increasingly willing to share these raw data. Yet there is too little awareness that the context of the experiments (lab notebook, sample preparation techniques, sensor configurations, etc.) also needs to be FAIR and shared to allow other researchers to truly understand the data, assess their usefulness (relevance, quality) for their purpose and for reproducibility, for example. A cultural change is required to convince researchers to offer this kind of contextual knowledge which is still seen as private information.

In general, we can also observe that there is an increase of awareness about the relevant legal and ethical regulations and state-of-the-art licensing mechanisms. Maintaining deep knowledge about all legal and ethical regulations is a challenge and requires a substantial effort. The same holds for issues such as licensing where researchers need to make choices from different options or need to exactly know what is allowed. Building up and maintaining the required knowledge on all regulations is adding considerable load on the researchers working with data.

2.2 Broken Data Cycle

In many cases we can observe that data generation and steps in data processing are carried out at different organizations using different methods and standards, i.e., the smooth data cycle as indicated in Figure 2 is broken. Data generators still deliver traditional data files, limited metadata in poorly documented form, if at all, and shift the responsibility for Openness[7] and FAIRness to the researchers who are then processing the data in their departmental environments. Currently, these are operating far away from being FAIR since manual operations, ad hoc scripts without documentation and the use of departmental or personal servers for generating and storing derived data are still the preferred practices. FAIR requirements such as assignment of PIDs and creation of rich comprehensive metadata including provenance information are not achieved. These non-FAIR practices are causing huge inefficiencies.

Some suggest storing final data related with "classical publications" ignoring that this is very inefficient, and it will only be carried out for a limited amount of data and will only provide the needed rich context if large curation efforts will be taken at the project end. Nonetheless, linking data sets to relevant publications does provide additional context although generally it is not machine actionable. Others suggest shifting the responsibility for the FAIRness of the generated data to the data creation labs. This could be a useful step, but not without closing the circle and changing the data practices in the data labs where the crucial processing producing results takes place.

[6] With machine actionability we refer to the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention.

[7] "Openness" in this note is always meant in the sense of "as open as possible".

2.3 Interdisciplinarity & Integration

Most data labs are still focusing on their immediate needs emerging from discipline specific research questions and are hardly thinking in terms of usability beyond their own narrow boundaries. Fostering interdisciplinary research requires some altruism since sufficient contextual information needs to be associated with data. Yet, there is little pressure to change practices, since data-driven cross-disciplinary research is still in its infancy. One consequence of this silo mentality is that many believe that their data requirements are special although the patterns according to which data management, curation (with provenance) and processing is organized are remarkably similar.

In general, research infrastructures are essentially integrating different centers (nodes) and offering joint services. Harmonizing standards and building integrative frameworks at different levels (identification, syntax, semantics, and portals) is not at all trivial; however, they have been successfully designed and developed in many different cases such as in the ESFRI initiatives. First, research infrastructures need to have these ambitions to build an umbrella and improve FAIRness and second, they should not re-invent the wheel but make use of existing knowledge and components.

2.4 Repositories

It is widely agreed that trustworthy and FAIR compliant repositories are the pillars of a stable and efficient data landscape [10]. Practices show, however, that most data are still being managed in centers where established practices are continued and where a large variety of software stacks (files, clouds, and databases) is being used. Internationally accepted assessments for trustworthiness such as CoreTrustSeal [11] are still not yet applied. Only stricter practices will finally help improving data management. In surveys people refer to different types of repositories for different data types (digital books/papers, data, metadata, etc.) which is an unnecessary complication at basic data management level. "Digital Object" is an abstract concept that includes all data types, i.e., repositories managing DOs are widely independent of the specific types.

2.5 Conclusions

From analyzing practices, we can draw a few conclusions:

- Compared with the findings from surveys within RDA in 2014 we can observe much more awareness about principles that should be followed but no significant changes in data practices in most data labs. There is also a process of terminology awareness and thus an implicit harmonization.
- Despite all progress, we see large differences in knowledge especially when it comes to the details of, for example, the FAIR principles. Integrating essential aspects such as assigning PIDs and associating rich metadata from the beginning are not supported in the daily work. Much more training and support is required to broaden deep understanding and much better software is required to support individual researchers and not to put further load on them.

- The "data publication" step has been improved continuously encouraging some to suggest applying the same methods in the data labs. Due to completely different requirements between DS and data publication, this would be doomed to fail.
- Data centers (repositories) need to take a much more active role in making steps towards certification and FAIRness and guiding their customers—the individual researchers.

3. PRINCIPLES AND BASICS

To overcome the challenges raised in Section 2 it is necessary to use appropriate technologies to support the researchers in (a) providing FAIR information and (b) using FAIR information for the purposes of research leading to improved wealth creation and quality of life. Alongside the provision of appropriate technologies, there is the need to ensure the end-user works within a well-governed environment covering topics as explained, for example, in Responsible Research and Innovation (RRI) [12].

3.1 Key Technologies

The key technologies for FAIR research are identification and metadata and both need to be supported by standardization, harmonization and a complex system of software services to enable their creation, conversion, editing and supplementing and the use of the metadata for the active verbs of FAIR. In this realm the concept of FAIR DOs is important due its abstraction, binding, and encapsulation potential.

Another important technology is workflow; to be able to support the steps in the research process is required from idea to observation or experiment to data collection and analysis to data curation and research findings publication. Within this context there are workflows for individual research steps and across steps.

3.2 Identification

It is commonly accepted that persistent identifiers (PIDs) will be a crucial component of our digital scientific memory for the coming centuries. It is not only the DOs that need to be stored, but also many types of relationships among the DOs—be they created manually or automatically. Therefore, people are starting to understand that identifiers as the basis for references need to be persistent over centuries if we do not want to lose scientific knowledge. It is also agreed that PIDs need to be universally unique and resolvable to useful machine actionable information about the DO. The harmonization of attributes associated with PIDs has been worked out in the RDA Kernel Group [13].

This need for identifiers lasting for very long periods and the need for a harmonized universal resolution system has convinced many data professionals that URIs as they are used in the ephemeral Web do not offer the persistence that is needed. Instead[8], many people are using Handles [14] or DOIs [15][9].

[8] Note that while most Handles currently are expressed on Web pages and other documents in the form of URIs, the underlying system sits directly on the core Internet protocols

[9] DOIs are Handles with a prefix 10 and currently used broadly in, but not restricted to, the domain of electronic publication.

3.3 Metadata

Commonly defined as "data about data" in reality it is not so simple. RDA Interest Group on metadata (MIG) [6] has defined an extended set of principles:

- The only difference between metadata and data is mode of use.
- Metadata is not just for data, it is also for users, software services, computing resources, thus for all kinds of DOs.
- Metadata is not just for description and discovery; it is also for contextualization (relevance, quality, restrictions (rights, costs)) and for coupling users, software, and computing resources to data (to provide a Virtual Research Environment). Scientific metadata is meant to support efficient DS.
- Metadata must be machine-understandable as well as human-understandable for autonomicity (formalism).
- Management (meta)data is also relevant (research proposal, funding, project information, research outputs, outcomes, impact…).

Metadata descriptions therefore are DOs and need to be FAIR compliant, i.e., they should be identified by a PID, should be findable by being harvested by portals using standard protocols and accessible. The interoperability and re-usability dimensions of FAIR require that all metadata schemas are registered and available at publicly available registries and that all concepts and vocabularies being used in metadata descriptions are defined and also registered in publicly available registries. The role of metadata between end-users and a DO of interest may best be expressed as a diagram (Figure 3).

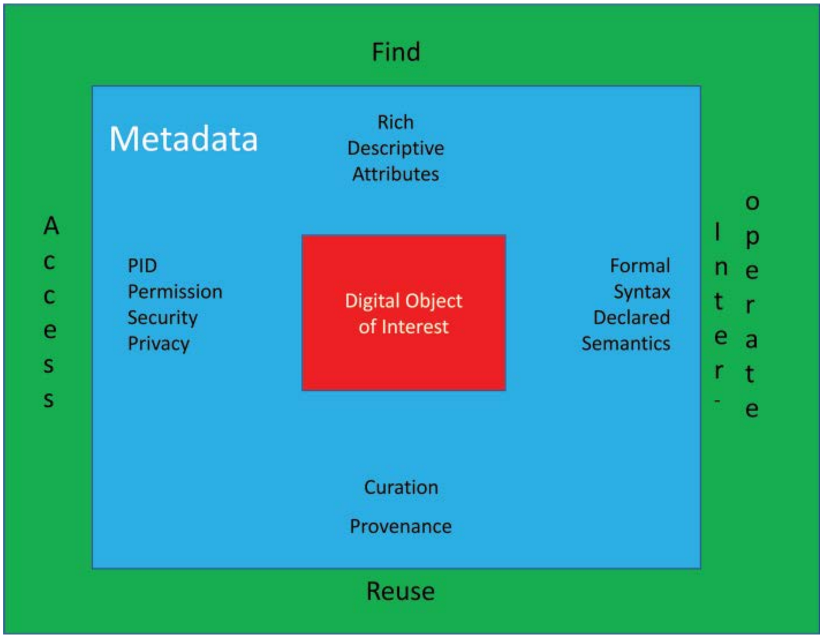

Figure 3. Metadata and FAIR.

Thus, the metadata exposes the DO to the world and simultaneously can help to protect it against unauthorized use. It provides rich attributes for discovery, controls for access, a formal syntax and declared semantics to encourage interoperation—including conversion between metadata formats—and information on curation and provenance for assessing relevance and quality and also for optimizing data locality.

There are many metadata *de facto* or *de jure* standards. The RDA Metadata Standards Catalog documents most of the popular ones [16]. Some are applicable to a small group of researchers allowing encoding of disciplinary specifics that are important for research; some are used very generally over a wide range of application domains and users. Inevitably, the general standards are good for finding digital assets over a wide range of domains, whereas specific standards for a narrow research area can—if they have the relevant properties (see Figure 3)—support each aspect of FAIR and highly automated DS. Since there is no single accepted metadata standard, inevitably we must convert between metadata standards to provide a homogeneous FAIR view over heterogeneous digital assets. If we convert from each metadata standard to all others, we need to develop *n(n-1)* two-way convertors (a.k.a. brokers). If, generally or even in one domain, we choose a single canonical (rich) metadata standard and convert the other relevant metadata standards to that, we have to develop only *n* convertors. The mappings to define the parameters for the convertors (usually as stored tables of relationships between entities and attributes in each of the two metadata standards) may be done manually. Tools exist to support this work such as for example the X3ML toolkit [17].

For many domains of research, as well as metadata to manage digital assets FAIRly, it is necessary to have additional rich metadata. Examples are measures of precision and accuracy. These may be considered rich attributes. Similarly, it may be necessary to understand the control parameters of an experiment or observation and any other relevant conditions that are stored, usually, in a lab notebook. This information may be considered part of provenance. This is contextual metadata which assists an end user in finding digital assets and assessing them for relevance and quality but also assists in understanding the context within which the research was done.

The cost of creating and maintaining metadata may be large. Efforts to automate metadata collection are underway, and the cost is considerably reduced if the metadata is collected incrementally (and ideally automatically from the workflow) as the workflow proceeds. Given the trend to an increased use of workflow frameworks to automatize processes it should be added that such processes require highly detailed and typed provenance metadata that even goes down to versions of software being used. Otherwise process-enabled orchestration is not possible.

3.4 FAIR DOs

DOs as they have been defined by RDA DFT [4] are autonomous first-class entities on the Internet, since they bind together all relevant information that is necessary to process them. A DO has a structured bit-sequence encoding its content which is stored and managed in some repositories, it has an assigned PID and is associated with rich metadata as described above. It is the Kernel information associated with the PID and yielded as result of the PID's resolution that contains important information and references in a persistent way. The Kernel Information types have been defined by the RDA Kernel Group [13] and have been registered in open registries [18]. The DO concept (Figure 4) supports abstraction in so far that at management level it does not make a difference between the type of the object (data, metadata, software, ontology, relation, etc.). Its encapsulation capability is given by associating a type (metadata attribute) with each DO and specifying operations that can be executed on this type. Also types and operations are identified by PIDs enabling the use of a fairly simple and straightforward interface protocol which allows interaction with DOs independently of the repository software and organization used [19].

Recently, it was suggested that, in addition to type registries, type ontologies should become available over time and that requirements for the machine actionability of metadata descriptions as mentioned above are required to speak about FAIR DOs [5].

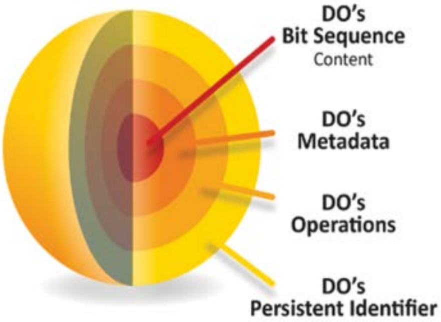

Figure 4. The DO concept.

3.5 Workflow

As mentioned earlier, research processes can be envisioned as workflows. The technology required is to support the process using appropriate IT, but purely IT-driven workflow frameworks are not sufficient, as practices show. Metadata provides a solid base for representing the digital assets and the processes acting on them with associated security, privacy, and monitoring. In a typical researcher workbench environment supported by IT (a.k.a. a virtual research environment (VRE)) the generic core process is something like this: (a) discover relevant digital assets; (b) assess them for relevance and quality; (c) compose a workflow using DOs (typically data and software services); (d) deploy the workflow onto appropriate computing/ storage/networking resources.

Steps (a) and (b) require the digital assets to be described by rich metadata so that they can not only be discovered and contextualized but also made interoperable by harmonizing the metadata descriptions, where possible, and converting the assets as necessary. Step (c) may be automated if the metadata is sufficiently rich, but often the researchers need to be involved in the composition. Current practice is that the workflow may be described in a canonical language such as Common Workflow Language (CWL) [20]. This has the advantage that the composition may be made independent of the deployment (step (d)) since several workflow managers accept CWL as input. Ideally the deployment should be optimized: the PaaSage project [21] demonstrated optimization of resource usage utilizing containers (Docker [22]) configured using Kubernetes [23]. However, a major factor in deployment optimization is the location of data: locality is required for efficiency (mainly due to limitations in network bandwidth). The MELODIC project—now a commercial offering [24]—developed such a system.

However, it should be noted that the existence of well-designed technical workflow frameworks such as Kepler, Taverna, Kmime, etc. did not change data practices in the data labs. There is an urgent need to understand what is needed to convince researchers to use such self-documenting workflow frameworks, which have the inherent capability to create FAIR compliant DOs. As has been shown in Section 2 it is urgently needed to make practices much more efficient, but to also help the researchers to facilitate his tasks without asking for continuous help from IT experts.

4. THINGS URGENTLY TO BE DONE

4.1 Paradoxes

The previous chapters indicated that we are faced with some paradoxes and that successful actions can only be scheduled if we understand the sources of them.

- "Standards" are good for science, since they have the potential to reduce efforts and costs substantially. *But, as Strawn points out, "standards" are bad for the researchers since they would have to change their ways of doing which means a loss in efficiency at least temporarily.*
- We have excellent principles and concepts (FAIR, FAIR DOs, etc.) that are based on a deep understanding of what is missing.
  The principles and concepts get strong support from many data professionals; however, researchers hesitate to change their practices in the data labs if they do not see immediate gains.
- We have a plethora of largely excellent tools, solutions and services promising to improve data practices ranging from metadata to workflow and semantic support based on deep IT knowledge.
  Yet all these tools and services, which were created mostly on individualistic insights and interests, have not had much effect on the improvement of current practices to substantially reduce the inefficiencies. The availability of easy-to-use services such as Zenodo, B2Share, FigShare, etc., for typical deposit actions has convinced some to make use of them and can be seen as exceptions.
- DS imposes new regulations and loads on the researchers (legal and ethical aspects, licenses, DMP creation, metadata creation, semantic explicitness, etc.).
  Yet, researchers prefer not to change their practices, but shift all changes to either the beginning (static DMP) or the end of the processes, when a publication needs to be associated with offering a collection containing all relevant data and beyond.
- Most data (>> 90%) is being created, managed and reused in the data lab processes and only a small percentage (<<10 %) is part of classical publications.
  Nevertheless, many hope that Open Science (OS) can be achieved by relying on the final step of data publication. This is where librarians and publishers are focused and this is supported by the researchers since it does not require changing practices.
- Discipline experts believe that their data challenges are unique.

Deep insights across disciplines, however, indicate that there are recurring patterns, but hardly anyone dares to touch this issue sensing the non-trivial challenges to develop cross-disciplinary tools. Cloud systems indicate, however, at a conceptually simple level that data commons are possible.

- Many speak about the opportunities of interdisciplinary research.
  Most often, however, data management efforts in the data labs are directed to the needs of a narrow research community indicating that support for cross-disciplinarity is in its infancy.
- It is now widely known that data projects are suffering from needing to spend about 80% of available time on mundane data wrangling.
  Yet, the big labs that can participate in DS seem to have the funding support to cover these inefficiencies and thus do not see the need to improve.

The main paradoxes can be condensed to the tension between two states as follows:

- narrow disciplinary *versus* broad interdisciplinary efforts;
- the start of the scientific process (data collection) *versus* the end of the scientific process (publication);
- conservative behaviors (stay with what you know) *versus* radical behavior (change your way of working);
- funding of conservative behaviors *versus* difficulty of obtaining funding for radical behaviors (conservatism of reviewers/funders); and
- independent activity of individual researchers *versus* various regulations and conditions.

These paradoxes indicate that despite paper declarations we did not yet manage

- to substantially change the practices in the thousands of data labs;
- to convince the researchers to apply new methods and tools; and
- to build software frameworks that can help researchers to effectively transition to efficient DS and allow them to carry out their work relatively independent of sparse IT specialists.

Of course, there are projects that are working on changes of practices towards more FAIRness and openness, but these examples do not seem to have affected the large mass of data labs. Cited examples often refer to the final step of "publishing data". The DRIHM community, for example, moved four years ago to recommend storing meteorology and climate data in B2SHARE [25]. They have formal agreements with EUDAT to be able to act as a community and thus could design their own metadata set with specific attributes relevant for the domain. The B2SHARE service is responsible for generating a Handle and a DOI for each uploaded set, to offer a portal to allow finding all stored data, and an API that allows harvesting metadata. This type of usage is thus an intermediary form in so far as it is not directly coupled with the process but does not require waiting on some future step where data may be published together with a classical paper.

Another example is the VODAN project [26] established in the realm of the GO FAIR initiative to help in COVID research. In this project a set of FAIR-Data-Points (FAIR compliant repositories) is being created that can immediately interchange metadata descriptions based on a common metadata model using well-defined and registered attributes. This as an excellent example of how FAIR compliance and a high degree of openness can be achieved easily, which is especially important for the health domain in which fragmentation is extreme. However, the project does not yet offer a model that includes the huge databases already aggregated in most disciplines.

Many other positive examples could be mentioned, but we need to admit that we are far away from having solutions that affect the broad scope of research actions. This is not surprising since the step from concepts and individual tools to universally applicable infrastructures functioning in a stable way takes time, as the Internet and the Web have shown.

4.2 Key Messages

Attempts to overcome the obvious roadblocks hidden in the described paradoxes are not simple and primarily not technical. Some key messages need to be considered before success can be achieved.

In general, researchers are not genuinely interested in new IT concepts and tools such as FAIR, FAIR DOs and FAIR implementations. These only become attractive when researchers see or expect a clear benefit for their research and when these new methods will be available in a stable way supported by easy-to-use software. All changes must be accompanied with educational efforts, which take time, and researchers will only invest if there is a chance of stable provisioning. The training effort is inversely correlated with the affinity of the known terminology and practices with those the researchers are familiar with, i.e., making use of appropriate branding is important.

There seems to be an increased awareness that the changes towards a FAIR data landscape will require quantum leaps which will imply a phase of instability and considerable learning efforts. Therefore, the key message for making progress given the paradoxes is not to speak about technical details but to focus on reference implementations of frameworks that facilitate researchers' current work and meet researchers' current practices as widely as possible without giving up the goals but instead implementing them in a way that researchers are not confronted with implementation technicalities and loaded with more work.

Such frameworks need to focus on operations which are repeated in researchers' normal processes. This includes different aspects:

- Generating and aggregating incrementally rich documentation (metadata and contextual information) during the steps of a process and generating snapshots after every step, allowing researchers to re-start a process at any point without the need to re-enter existing information. These snapshots could also be used to transfer states to other institutes and colleagues for various purposes such as checks and replication.
- The process documentation needs to be of a sort that allows editing of specific steps, i.e., parameters could be quickly changed enabling new runs without having to re-enter all remaining aspects. This often includes steps such as checking for the right legal and ethical regulations to be followed, building collections, selecting subjects, etc.
- The documentation and result generation will be done in a way allowing FAIR compliance without putting extra burdens on the researchers' shoulders, i.e., FAIR DOs will be immediately created and

uploaded into repositories using institutional, departmental or project settings[9]. Since the framework is meant to be cross-disciplinary different schemas for metadata, for example, must be supported by invoking suitable editors.

- The framework should help the researchers more easily cope with the increasing number of regulations and best practices of different sorts relevant to data-driven work—be they legal, ethical, or technical.
- Systematic snapshots and comprehensive documentation are key for cross-departmental and cross-disciplinary reuse where possible, again without adding load to the researchers.
- This approach to processing documentation will reduce the effort of creation of DMPs, which are seen as bureaucratic acts, by the documentation created by scientifically productive workflows which will take off load from the researchers.

We should not ignore two psychological aspects. (1) Researchers like to be as independent as possible from experts external to their realm of control, in order to be flexible and fast. This is of great importance since many aspects and steps of research are not predictable. This implies that any framework that makes them dependent for repeating operations on experts will only be accepted as a temporary solution. (2) The mass of researchers will be sceptical of introducing rigid automated actions since scientific work is highly nonlinear, i.e., linearity of operations and repetitions do not fully describe the scientific process.

There are ongoing efforts in a variety of institutions and projects to create workflow fragments to automate highly repetitive operations, i.e., making use of workflows is not new in science. Yet these attempts are scattered and have not been setup to create a FAIR compliant landscape of DOs. It is time to look at many of these fragment solutions and understand the degree of flexibility built into them, the challenges they address and their limitations.

4.3 Actions

There are still many areas that require efforts to help improve current practices. These include investing in training and education, and in increasing awareness, making metadata machine actionable, and providing improved frameworks to allow researchers to do semantic crosswalks. Here we will focus on an effort that seems to be urgent and has not yet been addressed sufficiently, as the previous chapters have shown: **flexible canonical workflow frameworks for science (CWFS)**. We see CWFS as a layer on top of technical workflow frameworks that could be used to implement CWFS and so we do not reinvent the wheel. Technical frameworks such as the Common Workflow Language and easy-to-use tools such as Jupyter give hope that such implementations will be possible. It will also be possible to make use of other technical developments such as Research Objects defining containers for easy portability. But these technical solutions are not of primary relevance at this stage.

Instead, we suggest a major and coordinated effort that will focus primarily on scientific needs as they emerge from practices in the data labs, with the goal of developing CWFS and with the expectation that such a framework will be accepted by researchers, will help change the practices in the labs, will increase

[9] To limit the scope of this work we need to assume that all schemas and concepts being used have been defined and registered in registries, all of which is important to ensure machine-actionability.

efficiency in data projects, and finally help to come a FAIR domain of DOs. There is no doubt that the engagement of many young researchers in the various labs across disciplines will be required to achieve these goals. We need to turn the focus from looking at the end products of data-driven science to the process. Finding solutions will not be simple, but we need to start taking steps.

With interested and ambitious pilot communities, we need to study recurring patterns in the preparation and execution of data-driven projects. As indicated, we need to analyze the experiences from ongoing work in automating process fragments. As an example, we mention a recent analysis that has been carried out in some research disciplines running experiments with humans which resulted in the pattern shown in Figure 5. The diagram shows a set of canonical steps that must be taken for every experiment and thus can be automated widely. The support for researchers needs to start with the hypothesis and will end with the registration and upload of the experimental data to a trustworthy repository of the researchers' choices. Not all the atomic steps will be necessary for all experiments, but those not needed could be simply ignored. Of course, for each step there might be differences per discipline, i.e., for each step there could be a library of packages the researchers could select from. Some steps imply complex nested workflows such as the experimental setup which includes a selection of many human subjects that are ready to participate in the experiment and as the execution step which in general includes repetitions over a few dimensions such as experimental items and human subjects.

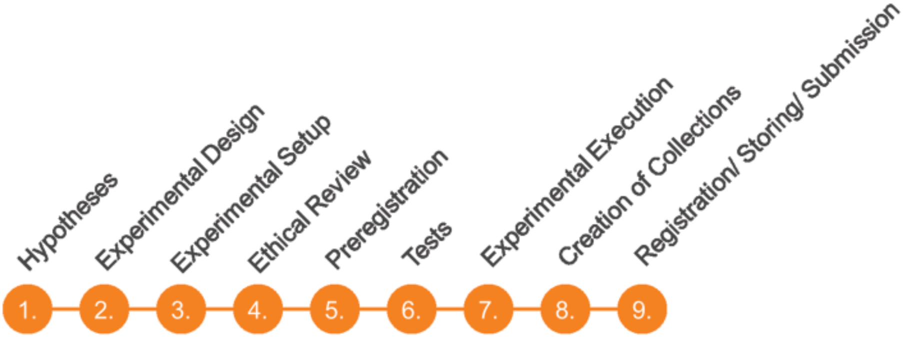

Figure 5. Cumulative comprehensive metadata.

We will not go into more detail in this paper, but want to point out that the creation of rich metadata and the inclusion of contextual information can be done stepwise, that the creation of DOs including the registration of persistent identifiers can be done automatically, that the upload in repositories can happen automatically, that support could be provided for following the required regulations, etc. It will be crucial to identify the interfaces that are required to allow the interaction between steps and to enable plugging in those packages that carry discipline-specific actions. Often, such specific actions are already managed by some smart software such as for example the selection of human subjects from a possibly big pool. Such software needs to be integrated into larger frameworks, which is a challenge.

Designing CWFS is not trivial, but the goal of helping to change data practices in the labs towards a FAIR compliant data domain while simultaneously saving time and money through increased efficiency makes it worth starting CWFS.

5. CONCLUSIONS

We began this paper describing the hopes for an imminent convergence of FAIR principles and DOs bringing about a FAIR compliant, unified, and persistent global data landscape and preventing the "dark digital age" [27]. We then looked to the reality of the processes in the many data labs across disciplines and countries, which are far from being FAIR, and described the principles which would help overcoming the huge existing inefficiencies and thus create effectively an OS domain. The paradoxes indicate that there are serious roadblocks hampering a fast transformation of the practices.

Basically, researchers see FAIR data as a request to change the practices with which they are familiar and to invest more time (in creating metadata, in registering suitable PIDs, in creating DOs to be stored in trustworthy repositories, etc.) without seeing the benefit for their work. Those researchers that are carrying out data-driven science do it either in a restricted fashion or they have sufficient funds to accept the large inefficiencies and the high rate of failures. Therefore, the underlying problem is simple to state: unless the reward for a researcher from doing 'best practice' is greater than the effort to accomplish 'best practice' there is no incentive to do so. Thus, we must increase reward and reduce effort.

Our observations overlap widely with many points mentioned in the European Open Science Cloud (EOSC) FAIR WG report "Seven Recommendations for Implementation of FAIR Practice" [28] when it refers to "technical gaps, FAIR misunderstandings, differences between disciplines, etc. It describes very well the social roadblocks and we seem to come to similar conclusions about the key messages. The USA National Academy of Sciences (NAS) consensus study report [29] covers broadly the same ideas. Creating FAIR DOs costs more but we would stress that we need to start from the beginning and not focus on the end result which the current mainstream is focusing on. Putting efforts on the end products will lead to a huge loss and will even be more expensive.

Stressing the "data fabric" idea, however, requires new and better tool support, support that offers researchers the way to FAIR data and effective OS without adding extra work. Indeed, we need to reduce the load on them from new regulations and the increasing amount of repetition. Administrative acts such as creating DMPs could be largely replaced by productive instruments. However, to make this happen we need to develop tools that address the data lab needs. Investments in flexible CWFS that automatically create FAIR DOs and upload them in trustworthy repositories are urgently required to make progress.

AUTHOR CONTRIBUTIONS

All authors contributed ideas, text and review comments in the production of the paper. G. Strawn (gostrawn@gmail.com) contributed much to Section 1. Sections 3 and 4 were collated dominantly by K. Jeffery (keith.jeffery@keithgjefferyconsultants.co.uk), and the others by P. Wittenburg (peter.wittenburg@mpcdf.mpg.de).

REFERENCES

[1] Wittenburg, P., Strawn, G.: Common patterns in revolutionary infrastructures and data. Available at: http://doi.org/10.23728/b2share.4e8ac36c0dd343da81fd9e83e72805a0. Accessed 6 January 2021

[2] Wilkinson, M., et al.: The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3, Article No. 160018 (2016)

[3] Kahn, R., Wilensky, R.: A framework for distributed digital object services. International Journal on Digital Libraries 6(2), 115–123 (2006)

[4] RDA DFT: DFT core terms and model. Available at: http://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318. Accessed 6 January 2021

[5] Paris GEDE Workshop on Moving Forward on Data Infrastructure Technology Convergence. Available at: https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects/Paris-FDO-workshop. Accessed 6 January 2021

[6] RDA MIG. Available at: https://www.rd-alliance.org/groups/metadata-ig.html. Accessed 6 January 2021

[7] Hetne, K., Wittenburg, P.: FAIR technology matrix—Phase 2 impressions. Available at: https://github.com/GEDE-RDA-Europe/GEDE/blob/master/Events/RDA%2014th%20Plenary/Matrix%20about%20technologies%20used%20by%20RIs_Background%20document.pdf. Accessed 6 January 2021

[8] Stehouwer, H., Wittenburg, P.: RDA Europe: Data practices analysis. Available at: http://hdl.handle.net/11304/6e1424cc-8927-11e4-ac7e-860aa0063d1f. Accessed 6 January 2021

[9] RDA Data Farbic IG. Available at: https://www.rd-alliance.org/group/data-fabric-ig.html. Accessed 6 January 2021

[10] Turning FAIR into reality—Final report and action plan from the European Commission expert group on FAIR data. Available at: https://op.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1/language-en/format-PDF/source-80611283. Accessed 6 January 2021

[11] Core Trust Seal. Available at: https://www.coretrustseal.org/. Accessed 6 January 2021

[12] Responsible research and innovation. Available at: https://en.wikipedia.org/wiki/Responsible_Research_and_Innovation. Accessed 6 January 2021

[13] RDA Kernel WG. Available at: https://www.rd-alliance.org/groups/pid-kernel-information-profile-management-wg. Accessed 6 January 2021

[14] Handles. Available at: https://www.handle.net/. Accessed 6 January 2021

[15] DOI. Available at: https://www.doi.org/. Accessed 6 January 2021

[16] RDA Metadata Standards Directory. Available at: https://rdamsc.bath.ac.uk/#:~:text=The%20RDA%20Metadata%20Standards%20Catalog,to%20help%20address%20infrastructure%20challenge. Accessed 6 January 2021

[17] X3ML Toolkit. Available at: https://www.ics.forth.gr/isl/x3ml-toolkit. Accessed 6 January 2021

[18] GWDG data type registry. Available at: https://os.helmholtz.de/fileadmin/user_upload/os.helmholtz.de/Workshops/rda_de_16_schwardmann.pdf. Accessed 6 January 2021

[19] DOIP. Available at: https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf. Accessed 6 January 2021

[20] CWL. Available at: https://www.commonwl.org/. Accessed 6 January 2021

[21] Paasage. Available at: https://paasage.ercim.eu/. Accessed 6 January 2021

[22] Docker. Available at: https://www.docker.com/. Accessed 6 January 2021

[23] Kubernetes. Available at: https://kubernetes.io/. Accessed 6 January 2021

[24] Melodic. Available at: https://melodic.cloud/. Accessed 6 January 2021

[25] B2SHARE. Available at: https://b2share.eudat.eu/. Accessed 6 January 2021

[26] VODAN. Available at: https://www.go-fair.org/implementation-networks/overview/vodan/. Accessed 6 January 2021

[27] Cerf, C.: How to prevent a digital dark age. Available at: https://www.theguardian.com/media-network/2015/may/29/googles-vint-cerf-prevent-digital-dark-age. Accessed 6 January 2021

[28] EOSC FAIR WG. Seven recommendations for implementation of FAIR practice—Draft for consultation. Available at: https://www.eoscsecretariat.eu/eosc-liaison-platform/post/seven-recommendations-implementation-fair-practice-draft-consultation. Accessed 6 January 2021

[29] NAS Consensus report. Available at: https://www.nap.edu/catalog/25116/open-science-by-design-realizing-a-vision-for-21st-century. Accessed 6 January 2021

AUTHOR BIOGRAPHY

Keith Jeffery is an independent consultant working on EPOS, ENVRIplus and ENVRIFAIR as well as on advanced CLOUD computing and Virtual research Environments. He is past IT Director at STFC with 360,000 users, 1,100 servers and 140 staff. Keith holds three honorary visiting professorships, a Fellow of the Geological Society and the British Computer Society, a Chartered Engineer & IT Professional and an Honorary Fellow of the Irish Computer Society. Keith is past President of ERCIM and euroCRIS, and serves on international expert groups, conference boards and assessment panels. He had advised government on IT. He chaired the EC Expert Groups on GRIDs and on CLOUD Computing.
ORCID: 0000-0003-4053-7825

Peter Wittenburg was Executive Director of Research Data Alliance (RDA) Europe, Member of RDA Technical Advisory Board, Scientific Coordinator of European Data Infrastructure (EUDAT) and Technical Director of the CLARIN and DOBES Research Infrastructures. He set up and led the Technical Group with about 35 experts at Max Planck Institute (MPI) for Psycholinguistics and then led the Language Archiving Group with about 25 experts. Since 2000 he has played leading roles in a variety of European (funded by the European Commission) and national projects (funded by MPS, DFG, BMBF, NWO) and ISO initiatives (ISO TC37/SC4). He won the Heinz Billing Award of the Max Planck Society (MPS) for the advancement of scientific computation in 2011 and received an honorary doctorate from University Tübingen in 2013.
ORCID: 0000-0003-3538-0106

Larry Lannom is Director of Information Services and Vice President at the Corporation for National Research Initiatives (CNRI), where he works with organizations in both the public and private sectors to develop experimental and pilot applications of advanced networking and information management technologies. Mr Lannom's current work is focused on CNRI's Digital Object Architecture, which is based on the concept of the DO, a uniform approach to representing digital information across computing and application environments, both now and into the future. Mr. Lannom joined CNRI in September of 1996. Prior to that, he was a Technical Director at DynCorp, Inc., where he served as an advisor on digital library research for the ISTO, CSTO, and ITO offices of the US Defense Advanced Research Projects Agency (DARPA), including initiating the Computer Science Technical Reports (CS-TR) project, DARPA's first effort in the digital library area. In addition, he managed the development of internal information systems for DARPA. Originally trained as a librarian, his earlier work included reference book publishing and information retrieval research.
ORCID: 0000-0003-1254-7604

George Strawn is currently the director of the Board on Research Data and Information at the National Academies of Sciences, Engineering, and Medicine where he focuses on OS and FAIR data. Prior to joining the Academies, Dr. Strawn was the director of the National Coordination Office (NCO) for the Networking and Information Technology Research and Development (NITRD) Program and co-chair of the NITRD interagency committee.
ORCID: 0000-0003-4098-0464

Claudia Biniossek works together with Dirk Betz on bringing together theories, methods, and data infrastructures within a transdisciplinary approach. The aim is to open new pathways of data-driven (transdisciplinary) research. Therefore, they created the repositories x-science.org and x-econ.org, which specialized in data coming from experimental social sciences and economics. The purpose is to test the basic principles of human decision making in the field of experimental data coming from economics, sociology, political sciences, psychology, and neuroscience.
ORCID: 0000-0002-2202-7875

Dirk Betz works together with Claudia Biniossek in bringing together theories, methods, and data infrastructures within a transdisciplinary approach. The aim is to open new pathways of data-driven (transdisciplinary) research. Therefore, they created the repositories x-science.org and x-econ.org, which specialized in data coming from experimental social sciences and economics. The purpose is to test the basic principles of human decision making in the field of experimental data coming from economics, sociology, political sciences, psychology, and neuroscience.
ORCID: 0000-0002-6411-4758

Christophe Blanchi is the Executive Director of the DONA Foundation in Geneva where he is responsible for its day-to-day operations, promoting and evolving the Digital Object Architecture and its related set of standards, and insuring the consistent operations of the Global Handle Registry. Prior to joining the DONA Foundation, Christophe Blanchi was a senior research scientist at CNRI in Reston, Virginia, where he was involved in research, development, and deployment of Digital Object Architecture related technologies.
ORCID: 0000-0003-2277-5176